# Aligning unequal multilingual thesauri

**Dan Ştefănescu**
Romanian Academy Institute for Artificial Intelligence
13, Calea „13 Septembrie”, 050711, Bucharest, Romania
E-mail: danstef@racai.ro

**Abstract**

The aligning and merging of ontologies with overlapping information are actual one of the most active domain of investigation in the Semantic Web community. Multilingual lexical ontologies thesauri are fundamental knowledge sources for most NLP projects addressing multilinguality. The alignment of multilingual lexical knowledge sources has various applications ranging from knowledge acquisition to semantic validation of interlingual equivalence of presumably the same meaning express in different languages. In this paper we present a general method for aligning ontologies which was used to align a conceptual thesaurus, lexicalized in 20 languages with a partial version of it lexicalized in Romanian. The objective of our work was to align the existing terms in the Romanian Eurovoc to the terms in the English Eurovoc and to automatically update the Romanian Eurovoc. The general formulation of the ontology alignment problem was set up along the lines established by Heterogeneity group of the KnowledgeWeb consortium, but the actual case study was motivated by the needs of a specific NLP project.

## 1. INTRODUCTION

The alignment of multilingual lexical knowledge sources has various applications ranging from knowledge acquisition to semantic validation of interlingual equivalence of presumably the same meaning express in different languages. In this paper we present a general method for aligning ontologies which was used to align a conceptual thesaurus, lexicalized in 20 languages with a partial version of it lexicalized in Romanian. The Romanian version of Eurovoc was incomplete not only because it misses one third of the terms but it also misses the cross-lingual unique identifiers. The objective of our work was to align the terms in the Romanian Eurovoc to the terms in the English Eurovoc and to automatically The general formulation of the ontology alignment problem was set up along the lines established by the Heterogeneity group of the KnowledgeWeb consortium (http://knowledgeweb.semanticweb.org/), but the actual case study was motivated by the needs of a national three-year research project called ROTEL. This project aims at the development of an integrated platform for semantically producing and processing large collections of multilingual documents (with the initial focus on Romanian-English language pair). The major multilingual data collection on which ROTEL platform will be tested is the 21-language AcquisCommunautaire (AC) parallel corpus (see Steinberger et al., 2006). The parallel documents are labelled with a common prefix, which is a CELEX unique identifier. The CELEX codes are on their turn associated with one or more EUROVOC concept codes. These associations (manually done) represent a gold-mine for any evaluation exercise for document classification and indexing engines.

The ROTEL platform will include several tools (some already existent, others to be developed) for multilingual language processing such as: language identification, tokenisation, POS tagging, chunking, dependency parsing, sentence alignment, word and phrase alignment, WSD, anaphora resolution, semantic annotation import, etc. There will be developed applications such as document classification, intelligent document indexing, document summarisation and question answering. For evaluation purposes (but not only), in the context of the AC corpus, the need for a Romanian version of the Eurovoc thesaurus is obvious. The Romanian version of the Eurovoc thesaurus is under development at the General Secretariat of the Chamber of Deputies of the Romanian Parliament. The only available document we could find about it was a PDF file with a two columns layout. Several terms, longer than a column, are partially shown and, in most cases, the unique term identifiers are not shown at all. Therefore, the task of recovering the Romanian version of Eurovoc, aligning it to the English hub version and importing the missing terms were challenging case studies for our ontology alignment platform.

Recently, the Eurovoc site[1] announced the release of the version 4.2, available in 17 languages, Romanian included. This is fortunate because we have now a gold-standard against which the ontology alignment system can be objectively evaluated.

## 2. EUROVOC

Eurovoc is a multilingual, polythematic thesaurus (Steinberger et al., 2002), which is used to index the Acquis Communautaire (the EU legislation and international treaties). Its fourth version is available in 20 languages out of which 16 are official EU languages.

The Romanian version of Eurovoc we used to validate the ontology alignment system was incomplete from multiple points of view:

a) it contains about 70% of the terms one could find in the English version;
b) the hierarchical structures are partial (there exist several dangling terms) and they are frequently different from the corresponding relations in the English version (it seems that the Romanian version follows the structuring of an early version of Eurovoc)

[1] http://europa.eu.int/celex/eurovoc

c) the cross-lingual unique identifiers, which allow the retrieving of the lexicalization of any Eurovoc term in any of the 20 languages, are not present in the Romanian version.

The Eurovoc thesaurus contains 6645 terms (519 top terms), covers 21 fields (from politics and international relations, to environment, industry or geography) and is structured into 127 microthesauri. The fields and microthesauri have unique identifiers in all languages allowing multilingual navigation. Each field is identified by a two-digit number while microthesauri are identified by four-digit numbers. The numbering of fields and microthesauri is the same for all languages. Each term is a node in one of the 519 trees rooted by the top terms. The Eurovoc contains five types of Semantic Relationships: *scope notes* (SN – definitions for clarifying the meaning of the descriptors), *microthesaurus relationships* (MT – references for descriptors showing their appurtenance to one or more microthesauri), *equivalence relationships* (UF, USE[2] – several types of relationships between descriptors and non-descriptors [3] ), *hierarchical relationships* (BT, NT – relationships between descriptors) and *associative relationships* (RT [4] – associative relationships between associated descriptors). These semantic relationships ensure the similarity between our problem and that of aligning ontologies and make us conjecture that any method which solves our problem can be employed for solving the aligning ontologies problem. The relevant relationships for our task were the hierarchical ones. The descriptors that do not have broader terms are called Top Terms.

## 3. THE TASK

Converting the PDF format of the Romanian version of Eurovoc into text format required recovering the trimmed out strings at the end of longer terms that didn't fit in the two column layout of the initial document.

| **Number of** | EN thesaurus | RO thesaurus |
|---|---|---|
| Descriptors | 6645 | 4625 |
| Top Terms | 519 | 508 |
| Reciprocal Hierarchical Relationships | 6669 | 3292 |
| Reciprocal Associative Relationships | 3636 | 2721 |

Table 1: Quantitative data for the EN and RO thesauri

Once this task completed, we constructed the hierarchical structures according to the specified relations and compared to the hierarchical structures of the English hub.

The Table 1 presents the quantitative data for the English and Romanian versions of the Eurovoc.

Our first goal was to align the existent terms in the Romanian version of Eurovoc to the English equivalents and this way to recover the terms unique IDs. Relying on the assumption that the structures in the two versions of Eurovoc should be identical, the next goal was to identify the missing terms and their respective relations. Generating translation equivalents for the identified missing terms was the last goal.

One should note that not having all the terms translated in the Romanian version made the problem harder to solve: the hierarchical relationships and the top terms ensure the existence of as many trees as the number of the top terms. In our case, we definitely had to expect that not all of the considered top terms in the Romanian version corresponded exactly to the top terms of the English version and that we would have to align incomplete tree structures, too.


Figure 1: Some Eurovoc Romanian trees

For the problem at hand, we consider that two or more trees have the same structure if they have the same structure of nodes disregarding the order of the sibling nodes in the tree.

## 4. SOLUTION

The thesauri alignment proceeds in two phases. The first one produces a backbone of the alignment, while the second one completes the alignment, identifying the missing terms and also producing suggestive raw translations for them.

The first phase of the thesauri alignment is a breadth-first partial matching algorithm for the trees contained in the two thesauri. Once the roots of two trees are successfully aligned, their respective sub-trees should also get aligned; otherwise the roots alignment should be reconsidered. The hard part of the algorithm is the identification of the most probable roots alignments. The data structures used

[2] UF = Used For – relationship between the descriptors and the non-descriptor(s); USE = $UF^{-1}$

[3] The several relationships types covered by UF and USE are genuine synonymy, near- synonymy, antonymy and inclusion

[4] RT = Related Term; the associative relationships can be of different types, from *cause and effect* to *location* or *characteristic feature*

in the recursion of the algorithm are two sorted lists $R_E$ and $R_R$ containing terms of the English and Romanian versions of the thesaurus. Initially, these lists contain the top terms in each language. Given that in both language versions the hierarchical relations are available, finding the top terms is almost a trivial task; however, due to incompleteness of the Romanian thesaurus one term might appear in the top term list just because its BT was not translated. Also, a non translated term may lead to the situation in which a single tree in English corresponds to more than one tree in Romanian. The shorter list ($R_R$ in our case) was appended with special symbols denoting empty translations. In order to identify the most probable term alignments, we used the COWAL aligner (Tufiş et al. 2005), trained on the Romanian-English sub-corpus of the Acquis Communautaire 21-languages parallel corpus. As expected, the translation model contains multiple statistical translation equivalents for (almost) any constituent word of an English descriptor. We used these translation equivalents for computing the most probable translations for each Romanian term.

```
double translation_score(string ro_text, string en_text) {
  Hashtable table;
  double ret;
  string[] ro_words = tokenize(ro_text);
  string[] en_words = tokenize(en_text);
  for (int i = 0; i < ro_words.Length; i++) {
    if (TE_prob[ro_words[i]].Keys.Count > 0) {
      foreach (key in TE_prob[ro_words[i]].Keys) {
        if (!table.ContainsKey(trans_equivalent[key]))
table.Add(trans_equivalent[key],TE_prob[ro_words[i]][key]);
          else
  if(table[trans_equivalent[key]] <TE_prob[ro_words[i]][key]
   table[trans_equivalent[key]] = TE_prob[ro_words[i]][key];
        }
      }
    }
  bool flag = false;
  for (int i = 0; i < en_words.Length; i++) {
    if (table.ContainsKey(en_words[i]) && table[en_words[i]]
> THRESHOLD)
      ret += table[en_words[i]];
    else
      flag = true;
  }
  if (flag)
    return 0;
  else
    return ret;
}
```

Figure 2: Algorithm for computing the translation score between two terms.

This is done using a translation score as it follows: for each word of a Romanian term, each English translation equivalent is introduced into a hash table (*TE_prob* in Figure 2) along with its estimated probability. If the equivalent is already in the hash table, then its estimated probability is updated with the greatest value between the old and the new one. In this same way, for all the Romanian terms, hash tables are constructed. After this, a multi-iterative process starts. For each Romanian term, and for each English term, we compute the translation score as the sum of the estimated probabilities (higher than a threshold), of the words which form the English term and are in the Romanian term hash table created above.

If an English word composing a term can not be found in the hash table or its value in the TE_prob table is lower than a certain threshold, then the score is nil.

In Figure 2 is shown the algorithm for computing the translation score between two terms.

The maximum translation score should indicate an English term as the translation of the Romanian term but all the other translation scores, greater than a threshold, along with their correspondent English terms are kept in order to solve possible duplications in the translation. The highest score indicates the most probable translation and the Ro-En terms pair, which corresponds to it, is kept as a correct Ro-En translation. The terms involved in this pair are eliminated both from the $R_R$ and $R_E$ list and also from the possible translations of the terms in the $R_R$ list. This process is repeated until the lists remain unchanged. Of course, we could use the same procedure for the entire lists of descriptors, but many of them, which are tied by a hierarchical relationship, are so similar that they considerably lower the accuracy of the alignment. On the other hand, the top terms, as non-related descriptors, are lexically very different. This ensures the premises for a high accuracy alignment. However, we have to take into account the possibility that for some English trees, only some sub-trees of descriptors were translated into Romanian and so, some Romanian top terms can not be aligned with the English top terms or are wrongly aligned. In this step, we successfully aligned 358 Romanian top terms.

The time and memory resources for the task described are not expensive as the number of the Romanian and English terms in $R_R$ and $R_E$ lists is small. We should note that the root items, in any ontology, are also to be found in small numbers as they should be the most general concepts, and so, the above stage would have worked as well if instead of thesauri we had had ontologies.

At the end of this phase, the few remaining terms in $R_R$ were not proper top terms. In the next step, the terms in $R_E$ are replaced by their immediate successors (NT) while the content of $R_R$ remains unchanged. The rationale is that the $R_R$ list terms might be aligned with one of the sub-trees of an English higher level term. This part is also repeated until no term remains in the $R_R$ list. In case of some terms still remain in the $R_R$ list, it is because that they are part of a sub-trees of some already aligned terms. These terms are kept in a special list and we should mention that their number is expected to be very small.

For every pair aligned in the entire process described above, we repeat the whole procedure. This time, the $R_E$ list is formed by the narrow terms of the English term in the pair, and the $R_R$ list by the narrow terms of the

Romanian term in the pair, plus the terms in the special list. If a term in the special list is found to correspond to an English term, it is removed from the special list. This also is repeated recursively until all Romanian terms find their English pairs or until the entire structure is parsed.
In case the entire structure is traversed but there still are Romanian terms unaligned, these terms are discarded as bad or wrong entries. In the end, we successfully aligned 4136 terms which means that 489 descriptors were discarded. The resources needed are kept significantly low because the work is gradually focused on hierarchical levels and also because our algorithm uses integers instead of strings.
The completion step takes care of the English descriptors that did not receive an index in the previous phase. The mapping tree-structure algorithm traverses the hierarchies of the two parallel thesauri and inserts dummy nodes in the Romanian thesaurus for the missing terms (not yet translated), in such a way as to preserve the English structure of the thesaurus. The translation model of COWAL is used to produce rough translations, indicative enough for the expert who usually is expected to edit it (the translation model is lemma based, and therefore a multiword term is translated as a sequence of lemmas) and to validate the proposed terms. A GUI interface allows the expert to visualise both Romanian and English thesauri, the aligned parts of them, and to edit the Romanian thesaurus for correction or for adding new information (such as multiple non-descriptors, not necessarily paralleled in the English version).

## 5. EVALUATION

Recently, we learnt about the existence of a Romanian version of the Eurovoc in its last release (version 4.2). We compared our reconstructed Eurovoc with the gold-standard version included into the last release. The first comparison concerned the mapping of the existing terms. The result (86.02%) was very disappointing and therefore, we analyzed the 576 differences to find out what was wrong in the alignment. We were happy to discover that **none of the differences** was a mapping error; the differences appeared because the terms in our version were revised in the version included into the Eurovoc 4.2 release. Therefore, we may say that the alignment was perfect. Table 3 exemplifies a few of the 576 Romanian terms that were reformulated in V4.2.
The second part of the evaluation refers to the proposed translation for the missing terms in the Ro1 version. Our investigation shows that 72% of the proposed term translations are correct[5].

| *ID* | *En term* | *Ro1 term* | *Ro4.2 term* |
|---|---|---|---|
| 15 | committee of inquiry | comisie de anchetă | comisie parlamentară de anchetă |
| 556 | housing law | legea locuinței | drept locativ |
| 983 | collective farm | grup de ferme | fermă colectivă |
| 1268 | nutrition | nutriție | alimentație |
| 1164 | financial management | gestiune financiară | management financiar |
| 3025 | political system | sistem politic | regim politic |
| 3179 | social rehabilitation | reabilitate socială | reabilitare socială |

Table 3: Examples of Romanian terms (Ro1) reformulated in the last Eurovoc release (Ro4.2)

## 6. CONCLUSION

Aligning multilingual thesauri is a very time-consuming and labour-intensive task when is manually done. We have presented a reasonable fast and very reliable method for automated aligning of such multilingual thesauri. Although the reported work was motivated by a very specific requirement, the system we developed is applicable to any other similarly structured thesaurus and is easy to extend/adapt for working with more elaborated hierarchical knowledge structures such as ontologies of the Semantic Web.

**Acknowledgements:**
We are grateful to the Office for Official Publications of the European Communities for granting us the Eurovoc licence. The work reported here is based on the original language editions of the *Eurovoc Thesaurus (Edition 4.2)* © European Communities, 2006. Responsibility for the adaptation lies entirely with the Research Institute for Artificial Intelligence.

## REFERENCES

Steinberger, R., Pouliquen, B., Hagman, J. (2002). Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus Eurovoc, Springer-Verlag.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C. Erjavec, T., Tufiş, D., Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages.

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2005). Combined word alignments, Proceedings of the ACL Workshop on Building and Using Parallel Texts (pp. 107—110), Ann Arbor.

[5] However, this figure might be significantly higher if one considers that the proposed translations were sequences of lemmas. A fair comparison would require lemmatization of all the Ro4.2 terms.